

Hyper- g Priors for Generalised Additive Model Selection with Penalised Splines

Daniel Sabanés Bové* Leonhard Held* Göran Kauermann†

Version: 18th August 2011

We propose an automatic Bayesian approach to the selection of covariates and their penalised splines transformations in generalised additive models. Specification of a default, hyper- g prior for the model parameters and a multiplicity-correction prior for the models themselves is crucial for this task. We introduce the methodology in the normal model and extend it to non-normal exponential families. Two applications from the literature illustrate the proposed approach. An efficient implementation is available in an R-package.

Keywords: penalised splines, Bayesian variable selection, g -prior, shrinkage.

1. Introduction

Semiparametric regression has achieved an impressive dissemination over the last years. Its central idea is to replace parametric regression functions by smooth, semi-parametric components. Following [Hastie and Tibshirani \(1990\)](#), suppose we have p

*Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland.

E-mail: {[daniel.sabanesbove](mailto:daniel.sabanesbove@ifspm.uzh.ch), [leonhard.held](mailto:leonhard.held@ifspm.uzh.ch)}@ifspm.uzh.ch

†Centre for Statistics, Department of Economics and Business Administration, University Bielefeld, Germany. E-mail: gkauermann@uni-bielefeld.de

continuous covariates x_1, \dots, x_p and use the additive model

$$y = \beta_0 + \sum_{j=1}^p m_j(x_j) + \epsilon, \quad (1)$$

where m_j are smooth but otherwise unspecified functions and $\epsilon \sim N(0, \sigma^2)$. For identifiability purposes we further assume that $\mathbb{E}(m_j(X_j)) = 0$ with respect to the marginal distribution of covariate X_j . Estimation of the smooth terms in (1) can be carried out in different ways, where we here make use of penalised splines, see *e.g.* [Eilers and Marx \(2010\)](#) or [Wood \(2006\)](#). A general introduction to penalised spline smoothing has been provided by [Ruppert, Wand, and Carroll \(2003\)](#) and the approach has become a popular smoothing technique since then, see [Ruppert, Wand, and Carroll \(2009\)](#). The general idea is to decompose the function m_j into a linear and a nonlinear part, where the latter is represented through a spline basis, that is

$$m_j(x_j) = x_j \beta_j + \mathbf{Z}_j(x_j)^T \mathbf{u}_j. \quad (2)$$

Here $\mathbf{Z}_j(x_j)$ is a $K \times 1$ spline basis vector at position x_j and \mathbf{u}_j is the corresponding coefficient vector. Conveniently one may choose a truncated polynomial basis for $\mathbf{Z}_j(\cdot)$ but representation (2) holds in general as well, see [Wand and Ormerod \(2008\)](#). To achieve a smooth fit one imposes a (quadratic) penalty on the spline coefficient vector \mathbf{u}_j which is formulated as the normal prior

$$\mathbf{u}_j \sim N_K(\mathbf{0}_K, \sigma^2 \rho_j \mathbf{I}_K), \quad (3)$$

where $\mathbf{0}_K$ is the all-zeros vector and \mathbf{I}_K is the identity matrix of dimension K . Here the variance factor ρ_j steers the amount of penalisation (relative to the regression variance σ^2). A larger ρ_j leads to a higher prior variance of the spline coefficients and hence a more wiggly function m_j , while a smaller ρ_j leads to a stronger penalty on $\|\mathbf{u}_j\|$ and thus a smoother function m_j . Setting ρ_j to zero imposes $\mathbf{u}_j \equiv \mathbf{0}_K$ so that $m_j(x_j)$ collapses to a linear term $m_j(x_j) = x_j \beta_j$. Hence the role of ρ_j ($j = 1, \dots, p$) extends to the selection of (generalised) additive models, which will be the focus of this paper. Variable selection will be treated by allowing the alternative $m_j(x_j) \equiv 0$.

Variable selection in generalised additive models is important to reduce the variance of effect estimates due to uninformative covariates. The field is wide and many

different approaches have been proposed in the last years. [Friedman \(2001\)](#) and [Tutz and Binder \(2006\)](#) describe boosting algorithms, which are extended by [Kneib, Hothorn, and Tutz \(2009\)](#) to geoadditive regression models ([Fahrmeir, Kneib, and Lang, 2004](#)). For the same model class, [Belitz and Lang \(2008\)](#) propose to use information-criteria or cross-validation, while [Fahrmeir, Kneib, and Konrath \(2010\)](#) and [Panagiotelis and Smith \(2008\)](#) use spike-and-slab priors for variable and function selection. [Brezger and Lang \(2008\)](#) adopt the concept of Bayesian contour probabilities ([Held, 2004](#)) to decide on the inclusion and form of covariate effects. [Cottet, Kohn, and Nott \(2008\)](#) generalise earlier work by [Yau, Kohn, and Wood \(2003\)](#) to Bayesian double-exponential regression models, which comprise generalised additive models as a special case. Shrinkage approaches are proposed by [Wood \(2011\)](#) and [Marra and Wood \(2011\)](#). [Zhang and Lin \(2006\)](#) use a lasso-type penalised likelihood approach, and [Ravikumar, Liu, Lafferty, and Wasserman \(2008\)](#) and [Meier, van de Geer, and Bühlmann \(2009\)](#) use penalties favouring both sparsity and smoothness of high-dimensional models. Likelihood-ratio testing methods are described by [Kauermann and Tutz \(2001\)](#) and [Cantoni and Hastie \(2002\)](#). This list mirrors the multitude as well as the variety of the different approaches and the enumeration is, of course, in no way exhaustive.

In this paper we propose a novel Bayesian variable and function selection approach based on (generalised) hyper- g priors. This type of prior for the parameters in the generalised additive model traces back to the g -prior in the linear model ([Zellner, 1986](#)). Its hyper-parameter g acts as an inverse relative prior sample size, and assigning it a hyper-prior can solve the information paradox ([Liang, Paulo, Molina, Clyde, and Berger, 2008](#), section 4.1) of the fixed- g case ([Berger and Pericchi, 2001](#), p. 148) in the linear model. We will subsequently refer to such mixtures of g -priors generically as hyper- g priors. One specific example is the hyper- g prior of [Liang et al. \(2008](#), section 3.2), which enjoys a closed form for the marginal likelihood and leads to consistent model selection and model-averaged prediction. We follow the conventional prior approach ([Berger and Pericchi, 2001](#), section 2.1) by using non-informative improper priors for parameters which are common to all models, and default proper hyper- g priors for model-specific parameters.

The current work generalises the hyper- g prior for generalised linear models ([Sabanés Bové and Held, 2011a](#)). In the same paper, we showed how fractional polynomials ([Sabanés Bové and Held, 2011b](#)), which extend ordinary polynomials by square roots, reciprocals and the log-

arithm, can be used to model nonlinear covariate effects. However, fractional polynomials have the disadvantage of being not invariant to linear transformations of the covariates. For variable and function selection, [Fahrmeir et al. \(2010\)](#) and [Scheipl \(2010\)](#) use a mixture of two inverse-gamma distributions with a very small (“spike”) and a larger mean (“slab”) as a hyper-prior for the variances of the regression coefficients’ independent normal priors. The posterior probability for inclusion of a coefficient is then estimated from the proportion of Markov chain Monte Carlo (MCMC) variance samples in the “slab”. While this prior structure eases the MCMC algorithm, it does not take into account the correlation structure of the covariates, and depends on the specification of the prior means in the two mixture components. [Cottet et al. \(2008\)](#) also use independent normal inverse-gamma priors for the regression coefficients, but they explicitly exclude coefficients from the model. For nonlinear effects they utilise low-rank approximations of smoothing splines, which require the choice of a threshold on the eigenvalue scale.

The paper is organised as follows. We first apply the hyper-g prior of [Liang et al. \(2008\)](#) to additive models in Section 2. The methodology is extended to generalised additive models in Section 3. A multiplicity-correction prior on the model space and a stochastic search procedure are proposed in Section 4. We apply our approach to real data in Section 5 and suggest postprocessing techniques in Section 6. Section 7 closes the paper with a discussion.

2. Hyper-g Priors for Additive Models

Assume we have observed independent responses y_i at covariate values x_{i1}, \dots, x_{ip} , $i = 1, \dots, n$, from the additive normal model (1). For each covariate $j = 1, \dots, p$, we stack the covariate values into the $n \times 1$ vector $\tilde{x}_j = (x_{1j}, \dots, x_{nj})^T$ and the spline basis vectors into the $n \times K$ matrix $\tilde{Z}_j = (Z_j(x_{1j})^T, \dots, Z_j(x_{nj})^T)^T$. The subsequent Gram-Schmidt process (see [Björck, 1967](#))

$$x_j = \tilde{x}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \tilde{x}_j}{\mathbf{1}_n^T \mathbf{1}_n} = \tilde{x}_j - \mathbf{1}_n \bar{x}_j, \quad (4)$$

$$Z_j = \tilde{Z}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \tilde{Z}_j}{\mathbf{1}_n^T \mathbf{1}_n} - x_j \frac{x_j^T \tilde{Z}_j}{x_j^T x_j}, \quad (5)$$

where $\mathbf{1}_n$ denotes the all-ones vector of dimension n , ensures that $\mathbf{1}_n$, x_j and the columns of \mathbf{Z}_j are orthogonal to each other, *i. e.* $\mathbf{1}_n^T x_j = 0$ and $\mathbf{1}_n^T \mathbf{Z}_j = x_j^T \mathbf{Z}_j = \mathbf{0}_K$.

A central measure of model complexity is the degrees of freedom. While in parametric models this is just the number of parameters, for smoothing and mixed models [Aerts, Claeskens, and Wand \(2002, section 2.2\)](#) translate the variance factor ρ_j into the corresponding degrees of freedom

$$d_j(\rho_j) = \text{tr}\{(\mathbf{Z}_j^T \mathbf{Z}_j + \rho_j^{-1} \mathbf{I})^{-1} \mathbf{Z}_j^T \mathbf{Z}_j\} + 1 \in (1, K + 1) \quad (6)$$

for a smoothly modelled covariate effect m_j . Note that $d_j(\rho_j) = \sum_{k=1}^K \lambda_{jk} / (\lambda_{jk} + \rho_j^{-1})$ is easy to calculate via the (positive) eigenvalues λ_{jk} of $\mathbf{Z}_j^T \mathbf{Z}_j$. This also shows that $d_j(\rho_j)$ is strictly increasing with derivative $\sum_{k=1}^K \lambda_{jk} / (\rho_j \lambda_{jk} + 1)^2 > 0$, which implies that we may (numerically) invert the function to $\rho_j(d_j)$. In fact, by fixing the degrees of freedom d_j for function $m_j(x_j)$ we define the variance factor ρ_j . Subsequently we will restrict the degrees of freedom to take values in a finite set $\mathcal{D} \subset \{0\} \cup [1, K + 1)$, say $\mathcal{D} = \{0, 1, 2, 3, \dots, K\}$. For $d_j = 0$ we set $m_j(x_j) \equiv 0$ while for $d_j = 1$ we have the linear model $m_j(x_j) = x_j \beta_j$. In general, model (1) is indexed by $\mathbf{d} = (d_1, \dots, d_p)$ giving the degrees of freedom for each functional component and hence the structure of the model.

After combining the $I = \sum_{j=1}^p \mathbb{I}(d_j \geq 1)$ vectors x_j to the $n \times I$ linear design matrix $\mathbf{X}_d = (x_j : d_j \geq 1)$ and the $J = \sum_{j=1}^p \mathbb{I}(d_j > 1)$ matrices \mathbf{Z}_j to the $n \times JK$ spline design matrix $\mathbf{Z}_d = (\mathbf{Z}_j : d_j > 1)$, and analogously constructing the respective coefficient vectors β_d and \mathbf{u}_d , the conditional additive model for the response vector $\mathbf{y} = (y_1, \dots, y_n)^T$ is

$$\mathbf{y} \mid \beta_0, \beta_d, \mathbf{u}_d, \sigma^2 \sim N_n(\mathbf{1}_n \beta_0 + \mathbf{X}_d \beta_d + \mathbf{Z}_d \mathbf{u}_d, \sigma^2 \mathbf{I}_n). \quad (7)$$

Integrating out the the spline coefficient vector $\mathbf{u}_d \sim N_{JK}(\mathbf{0}_{JK}, \sigma^2 \mathbf{D}_d)$, where \mathbf{D}_d is block-diagonal with J blocks $\rho_j \mathbf{I}_K$ ($d_j > 1$), yields the marginal model

$$\mathbf{y} \mid \beta_0, \beta_d, \sigma^2 \sim N_n(\mathbf{1}_n \beta_0 + \mathbf{X}_d \beta_d, \sigma^2 \mathbf{V}_d) \quad (8)$$

with $\mathbf{V}_d = \mathbf{I}_n + \mathbf{Z}_d \mathbf{D}_d \mathbf{Z}_d^T$. This general linear model can be decorrelated into a standard linear model by using the Cholesky decomposition $\mathbf{V}_d = \mathbf{V}_d^{T/2} \mathbf{V}_d^{1/2}$: For the transformed response vector $\tilde{\mathbf{y}} = \mathbf{V}_d^{-T/2} \mathbf{y}$ we have

$$\tilde{\mathbf{y}} \mid \beta_0, \beta_d, \sigma^2 \sim N_n(\tilde{\mathbf{1}}_n \beta_0 + \tilde{\mathbf{X}}_d \beta_d, \sigma^2 \mathbf{I}_n) \quad (9)$$

with analogously transformed all-ones vector $\tilde{\mathbf{1}}_n = \mathbf{V}_d^{-T/2} \mathbf{1}_n$ and design matrix $\tilde{\mathbf{X}}_d = \mathbf{V}_d^{-T/2} \mathbf{X}_d$. Note that now also $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{1}}_n$ depend on the model d , but we suppress this dependence for ease of notation.

We will now show how to use the hyper- g prior (Liang et al., 2008) for the parameters β_0 , β_d and σ^2 in the decorrelated model (9). The hyper- g prior consists of a locally uniform prior $f(\beta_0) \propto 1$ on the intercept, Jeffreys' prior $f(\sigma^2) \propto (\sigma^2)^{-1}$ on the regression variance and the g -prior (Zellner, 1986)

$$\beta_d | g, \sigma^2 \sim \mathbf{N}_I \left(\mathbf{0}_I, g\sigma^2 (\tilde{\mathbf{X}}_d^T \tilde{\mathbf{X}}_d)^{-1} \right) \quad (10)$$

on the linear coefficient vector, which is combined with a uniform hyper-prior on the shrinkage coefficient $g/(1+g)$. Note that the prior precision matrix in (10) is proportional to $\sigma^{-2} \tilde{\mathbf{X}}_d^T \tilde{\mathbf{X}}_d = \sigma^{-2} \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d$, which is the Fisher information matrix of β_d in model (8). Basically all formulae given by Liang et al. (2008) carry over to our setting, since inner products of the response vector \mathbf{y} , the all-ones vector $\mathbf{1}_n$ and the design matrix \mathbf{X}_d in model (8) carry over to their transformed counterparts $\tilde{\mathbf{y}}$, $\tilde{\mathbf{1}}_n$ and $\tilde{\mathbf{X}}_d$ in model (9). This is due to

$$\mathbf{V}_d^{-1} = (\mathbf{I}_n + \mathbf{Z}_d \mathbf{D}_d \mathbf{Z}_d^T)^{-1} = \mathbf{I}_n - \mathbf{Z}_d (\mathbf{Z}_d^T \mathbf{Z}_d + \mathbf{D}_d^{-1})^{-1} \mathbf{Z}_d^T, \quad (11)$$

which follows from the matrix inversion lemma (see Henderson and Searle, 1981) and leads to $\tilde{\mathbf{1}}_n^T \tilde{\mathbf{1}}_n = \mathbf{1}_n^T \mathbf{1}_n = n$, $\tilde{\mathbf{1}}_n^T \tilde{\mathbf{X}}_d = \mathbf{1}_n^T \mathbf{X}_d = \mathbf{0}_I$ and $\tilde{\mathbf{1}}_n^T \tilde{\mathbf{y}} = \mathbf{1}_n^T \mathbf{y}$ by straightforward calculations. A most convenient property of the hyper- g prior is that it yields a closed form marginal likelihood, which needs to be computed on the original response scale via the change of variables formula:

$$f(\mathbf{y} | d) \propto \left\| \mathbf{V}_d^{-T/2} (\mathbf{y} - \mathbf{1}_n \bar{y}) \right\|^{-(n-1)} (I+2)^{-1} {}_2F_1 \left(\frac{n-1}{2}; 1; \frac{I+4}{2}; \tilde{R}_d^2 \right) \left| \mathbf{V}_d^{1/2} \right|^{-1}, \quad (12)$$

where $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, ${}_2F_1$ is the Gaussian hypergeometric function (Abramowitz and Stegun, 1964, p. 558) and \tilde{R}_d^2 is the classical coefficient of determination in model (8) (see Appendix A.1 for implementation details).

Other hyper-priors could be assigned to g , but will typically not lead to a closed form of the marginal likelihood. Examples are the incomplete inverse-gamma prior on $1+g$ (Cui and George, 2008, p. 891), which generalises the above uniform prior on $g/(1+g)$, and an inverse-gamma prior on g , which corresponds to the Cauchy

prior of Zellner and Siow (1980). Liang et al. (2008) also describe a modified version of the hyper-g prior, called the hyper-g/ n prior, which is a special case of the conventional robust prior proposed by Forte (2011), for which a closed form of the marginal likelihood exists.

Posterior inference in a given model d is based on Monte Carlo estimation of the parameters in model (7), using the factorisation

$$f(\beta_0, \beta_d, \mathbf{u}_d, \sigma^2, g | \mathbf{y}) = f(\mathbf{u}_d | \beta_0, \beta_d, \sigma^2, \mathbf{y}) f(\beta_0, \beta_d | \sigma^2, g, \mathbf{y}) f(\sigma^2 | \mathbf{y}) f(g | \mathbf{y}). \quad (13)$$

Sampling of g, σ^2 and subsequently β_0, β_d can be done along the lines of Sabanés Bové and Held (2011b, section 2.3), by adapting their algorithm to the transformations in model (9). Finally, the spline coefficient vector \mathbf{u}_d is sampled from

$$\begin{aligned} f(\mathbf{u}_d | \beta_0, \beta_d, \sigma^2, \mathbf{y}) &\propto f(\mathbf{u}_d | \sigma^2) f(\mathbf{y} | \beta_0, \beta_d, \mathbf{u}_d, \sigma^2) \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \left[\mathbf{u}_d^T \mathbf{D}_d^{-1} \mathbf{u}_d + \|\mathbf{y} - \mathbf{1}_n \beta_0 - \mathbf{X}_d \beta_d - \mathbf{Z}_d \mathbf{u}_d\|^2 \right] \right\} \\ &\propto N_{JK} \left(\mathbf{u}_d | \Sigma_d \mathbf{Z}_d^T (\mathbf{y} - \mathbf{X}_d \beta_d), \sigma^2 \Sigma_d \right), \end{aligned} \quad (14)$$

where $\Sigma_d = (\mathbf{Z}_d^T \mathbf{Z}_d + \mathbf{D}_d^{-1})^{-1}$ and β_0 disappears because $\mathbf{Z}_d^T \mathbf{1}_n = \mathbf{0}_{JK}$.

Given posterior samples for the linear coefficient β_j and the spline coefficient vector \mathbf{u}_j for covariate j ($d_j > 1$), we would like to transform these into samples for the function $m_j(x_j)$, along a grid vector $\tilde{\mathbf{x}}_j^*$ of n^* points (on the same scale as the original $\tilde{\mathbf{x}}_j$ used for the model fitting). This is in principle straightforward, but one has to carefully apply analogous transformations as in (4) and (5) to $\tilde{\mathbf{x}}_j^*$ and the corresponding spline basis matrix $\tilde{\mathbf{Z}}_j^*$:

$$\mathbf{x}_j^* = \tilde{\mathbf{x}}_j^* - \mathbf{1}_{n^*} \frac{\mathbf{1}_n^T \tilde{\mathbf{x}}_j}{\mathbf{1}_n^T \mathbf{1}_n}, \quad (15)$$

$$\mathbf{Z}_j^* = \tilde{\mathbf{Z}}_j^* - \mathbf{1}_{n^*} \frac{\mathbf{1}_n^T \tilde{\mathbf{Z}}_j}{\mathbf{1}_n^T \mathbf{1}_n} - \mathbf{x}_j^* \frac{\mathbf{x}_j^T \tilde{\mathbf{Z}}_j}{\mathbf{x}_j^T \mathbf{x}_j}. \quad (16)$$

Afterwards, for each coefficient sample one can compute the corresponding vector of function values $m_j(\tilde{\mathbf{x}}_j^*) = \mathbf{x}_j^* \beta_j + \mathbf{Z}_j^* \mathbf{u}_j$. Similarly, prediction samples for the corresponding response vector \mathbf{y}^* can be extracted from the sampling output.

3. Hyper- g Priors for Generalised Additive Models

Now we extend the above setting and assume that the covariate effects $m_j(x_j)$ enter additively into the linear predictor

$$\eta = \beta_0 + \sum_{j=1}^p m_j(x_j) \quad (17)$$

of an exponential family distribution with canonical parameter θ , mean $\mathbb{E}(y) = h(\eta) = db(\theta)/d\theta$ and variance $\text{Var}(y) = \phi/w \cdot d^2b(\theta)/d\theta^2$ (see [McCullagh and Nelder, 1989](#)). We restrict our attention to non-normal distributions with fixed dispersion ϕ (as $\phi = 1$ for the Bernoulli and Poisson distribution) and known weight w . For n observations, the linear predictor vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ is

$$\boldsymbol{\eta} = \mathbf{1}_n \beta_0 + \mathbf{X}_d \boldsymbol{\beta}_d + \mathbf{Z}_d \mathbf{u}_d \quad (18)$$

and the likelihood is

$$f(\mathbf{y} \mid \beta_0, \boldsymbol{\beta}_d, \mathbf{u}_d) \propto \exp \left\{ \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{\phi/w_i} \right\}. \quad (19)$$

The main challenge for the derivation of a generalised g -prior is that the marginal density $f(\mathbf{y} \mid \beta_0, \boldsymbol{\beta}_d)$, which results from integrating out the spline coefficient vector

$$\mathbf{u}_d \sim N_{JK}(\mathbf{0}_{JK}, \mathbf{D}_d) \quad (20)$$

from (19), has no closed form. In particular, it is not Gaussian, in contrast to (8).

Before addressing this problem we first consider appropriate construction of the design matrices \mathbf{X}_d and \mathbf{Z}_d and calculation of the degrees of freedom $d_j(\rho_j)$ for a smoothly modelled term m_j . Starting with the latter, a reasonable generalisation of (6) is (see [Ruppert et al., 2003](#), section 11.4)

$$d_j(\rho_j) = \text{tr}\{(\mathbf{Z}_j^T \widehat{\mathbf{W}} \mathbf{Z}_j + \rho_j^{-1} \mathbf{I})^{-1} \mathbf{Z}_j^T \widehat{\mathbf{W}} \mathbf{Z}_j\} + 1, \quad (21)$$

which uses a fixed weight matrix $\widehat{\mathbf{W}} = \mathbf{W}(\mathbf{1}_n \widehat{\beta}_0)$, where $\mathbf{W}(\boldsymbol{\eta}) = \text{diag}\{(dh(\eta_i)/d\eta)^2 / \text{Var}(y_i)\}_{i=1}^n$ is the usual generalised linear model weight matrix and $\widehat{\beta}_0$ is the intercept estimate from the null model $\mathbf{d} = \mathbf{0}_p$. This definition avoids dependence of $\rho_j(d_j)$ on the model

d under consideration. As a consequence, we need to generalise the orthogonalisation of the original covariate vector \tilde{x}_j and spline basis matrix \tilde{Z}_j from (4) and (5) to

$$x_j = \tilde{x}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \widehat{W} \tilde{x}_j}{\mathbf{1}_n^T \widehat{W} \mathbf{1}_n} \quad (22)$$

$$\text{and } Z_j = \tilde{Z}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \widehat{W} \tilde{Z}_j}{\mathbf{1}_n^T \widehat{W} \mathbf{1}_n} - x_j \frac{x_j^T \widehat{W} \tilde{Z}_j}{x_j^T \widehat{W} x_j}, \quad (23)$$

implying that $\mathbf{1}_n$, x_j and the columns of Z_j are orthogonal to each other with respect to the inner product in terms of \widehat{W} . This ensures that (21) correctly captures only the degrees of freedom associated with the nonlinear part of m_j . Note that (15) and (16) are adapted analogously.

We will now derive a generalised g -prior analogous to (10) for the linear coefficient vector β_d in the generalised additive model. The idea is to use the iterative weighted least squares (IWLS) approximation to (19) to obtain an approximate normal model of the form (7) and then derive the resulting g -prior (10). With a slight abuse of notation, *e.g.* $h(\eta) = (h(\eta_1), \dots, h(\eta_n))^T$, let

$$z_0 = \eta_0 + \text{diag}\{dh(\eta_0)/d\eta\}^{-1}(\mathbf{y} - h(\eta_0)) \quad (24)$$

be the adjusted response vector resulting from a first-order approximation to $h^{-1}(\mathbf{y})$ around $\mathbf{y} = h(\eta_0)$. Then

$$z_0 | \beta_0, \beta_d, \mathbf{u}_d \stackrel{a}{\sim} N(\mathbf{1}_n \beta_0 + \mathbf{X}_d \beta_d + \mathbf{Z}_d \mathbf{u}_d, W_0^{-1}) \quad (25)$$

with $W_0 = W(\eta_0)$ is the working normal model (see *e.g.* McCullagh and Nelder, 1989, p. 40). Remember that the IWLS algorithm iteratively updates η_0 by weighted least squares estimation of the coefficients in (25). Here, we fix $\eta_0 = \mathbf{0}_n$, which is the value expected *a priori*. Then we rewrite (25) using $\tilde{z}_0 = W_0^{1/2} z_0$ etc. as

$$\tilde{z}_0 | \beta_0, \beta_d, \mathbf{u}_d \stackrel{a}{\sim} N(\tilde{\mathbf{1}}_n \beta_0 + \tilde{\mathbf{X}}_d \beta_d + \tilde{\mathbf{Z}}_d \mathbf{u}_d, I_n), \quad (26)$$

which brings us back to a normal model of the form in (7). By computing the corresponding g -prior (10), we arrive at the generalised g -prior

$$\beta_d | g \sim N_I(\mathbf{0}_I, gJ_0^{-1}) \quad (27)$$

with prior precision matrix proportional to

$$\begin{aligned} J_0 &= \tilde{X}_d^T (I_n + \tilde{Z}_d D_d \tilde{Z}_d^T)^{-1} \tilde{X}_d \\ &= X_d^T W_0^{1/2} (I_n + W_0^{1/2} Z_d D_d Z_d^T W_0^{1/2})^{-1} W_0^{1/2} X_d. \end{aligned} \quad (28)$$

An appealing feature of this prior is that it directly generalises the g -prior proposed by [Sabanés Bové and Held \(2011a\)](#) for generalised linear models, to which it reduces when there are no spline effects in the model, *i. e.* $J_0 = X_d^T W_0 X_d$. An alternative and more rigorous derivation of (28) as the Fisher information obtained from a Laplace approximation to the marginal model $f(\mathbf{y} | \beta_0, \beta_d)$ is presented in [Appendix B](#).

The generalised hyper- g prior

$$f(\beta_0, \beta_d, \mathbf{u}_d, g) = f(\beta_0) f(\beta_d | g) f(g) f(\mathbf{u}_d) \quad (29)$$

is defined to comprise the locally uniform prior $f(\beta_0) \propto 1$ on the intercept β_0 , the generalised g -prior (27) on the linear coefficient vector β_d , the penalty prior (20) on the spline coefficient vector \mathbf{u}_d , and some proper hyper-prior $f(g)$ on the hyper-parameter g . Posterior inference under this prior can be implemented by a straightforward extension of the approach of [Sabanés Bové and Held \(2011a, section 3\)](#), which is outlined in the following. The efficient R-package “hypergsplines” for this and all other computations in this paper is available from R-Forge.¹

Let $X_a = (\mathbf{1}_n, X_d, Z_d)$ and $\beta_a = (\beta_0, \beta_d^T, \mathbf{u}_d^T)^T$ denote the grand design matrix and regression coefficient vector, respectively, such that $\boldsymbol{\eta} = X_a \beta_a$. The prior for β_a conditional on g has a Gaussian form with mean zero and singular precision matrix $\text{diag}\{0, g^{-1} J_0, D_d^{-1}\}$. Thus, the Gaussian approximation of $f(\beta_a | \mathbf{y}, g, \mathbf{d})$, which is necessary for the Laplace approximation of $f(\mathbf{y} | g, \mathbf{d})$, can be obtained by the Bayesian IWLS algorithm ([West, 1985](#)). Afterwards, an approximation of the marginal likelihood of model \mathbf{d} ,

$$f(\mathbf{y} | \mathbf{d}) = \int_0^\infty f(\mathbf{y} | g, \mathbf{d}) f(g) dg, \quad (30)$$

is obtained by numerical integration of the Laplace approximation $\tilde{f}(\mathbf{y} | g, \mathbf{d})$. Note that recently integrated Laplace approximations have successfully been applied in a

¹The website is <http://r-forge.r-project.org/projects/hypergsplines>. To install the R-package, just type `install.packages("hypergsplines", repos="http://r-forge.r-project.org")` into R.

more general context (Rue, Martino, and Chopin, 2009). Finally, we can use a tuning-free Metropolis-Hastings algorithm to sample from the joint posterior of β_a and g in a specific model d .

4. Model Prior and Stochastic Search

We propose a prior $f(d)$ on the model space \mathcal{D}^p which explicitly corrects for the multiplicity of testing inherent in the simultaneous analysis of the p covariates (see Scott and Berger, 2010): *A priori*, the number of covariates included in the model (I) is uniformly distributed on $\{0, 1, \dots, p\}$. The choice of the I covariates is then uniformly distributed on all possible configurations, and their degrees of freedom are independent and uniformly distributed on $\mathcal{D} \setminus \{0\} = \{1, 2, 3, \dots, K\}$. Altogether, this gives

$$1/f(d) = (p+1) \binom{p}{I} K^I. \quad (31)$$

A nice property of this prior is that it leads to marginal prior probabilities $\mathbb{P}(d_j = 0) = \mathbb{P}(d_j > 0) = 1/2$. Elsewhere this is often achieved by assigning independent priors to the p covariates, which implies $I \sim \text{Bin}(p, 1/2)$ and has the disadvantage that models with $I \approx p/2$ included covariates are favoured by the prior. It is clear that our uniform prior on I allows the data y to have a maximum effect on the posterior of I because it is the reference prior (Bernardo, 1979).

Alternatively, one might also use a fixed (independent of K) prior probability for a linear effect ($d_j = 1$). This is appropriate for the situation where one explicitly wants to test linearity versus nonlinearity of each effect. Furthermore, a multiplicity correction for these tests can be implemented by assuming that the number of smoothly included covariates (J) is uniformly distributed on $\{0, 1, \dots, I\}$ and their choice is uniform on all possible choices. This would add one level to the prior hierarchy.

As the model space \mathcal{D}^p grows exponentially in the number of covariates p , only for small values of p all possible models can be evaluated. Otherwise the marginal likelihood $f(y|d)$ can be computed only for a subset of the model space. Usually this subset is determined by stochastic search procedures. Here we propose to use a simple Metropolis-Hastings algorithm with two possible move types in the proposal kernel:

Move Sample a covariate index $j \sim \text{U}\{1, 2, \dots, p\}$ and decrease or increase d_j to the next adjacent value in \mathcal{D} (with probability $1/2$ each, or deterministically if $d_j = 0$ or $d_j = K$, respectively).

Swap Sample a pair $(i, j) \sim \text{U}\{(1, 1), (1, 2), \dots, (p, p)\}$ of covariate indices ($i \leq j$) and swap d_i and d_j .

The ‘Swap’ move is designed to efficiently trace models with high posterior probability even in situations where covariates are almost collinear. For each Metropolis-Hastings iteration, a ‘Move’ is chosen with some fixed probability (we use $3/4$), and otherwise a ‘Swap’. Denote the current model by \mathbf{d} , then the proposed model \mathbf{d}' is accepted with probability

$$\alpha(\mathbf{d}' | \mathbf{d}) = 1 \wedge \frac{f(\mathbf{y} | \mathbf{d}')f(\mathbf{d}')q(\mathbf{d}' | \mathbf{d})}{f(\mathbf{y} | \mathbf{d})f(\mathbf{d})q(\mathbf{d} | \mathbf{d}')}$$

where the calculation of the proposal probability ratio $q(\mathbf{d}' | \mathbf{d})/q(\mathbf{d} | \mathbf{d}')$ is straightforward, see Appendix A.2.

The advantage of such an MCMC based model exploration compared to more elaborate stochastic search algorithms (*e.g.* [Hans, Dobra, and West, 2007](#); [Clyde, Ghosh, and Littman, 2011](#)) is that it does not preclude estimation of posterior probabilities via sampling frequencies, as it was originally proposed for MCMC model composition by [Madigan and York \(1995\)](#). Recently reported problems with renormalized probability estimates ([Clyde and Ghosh, 2010](#); [García-Donato and Martínez-Beneito, 2011](#)) can be avoided by using the model sampling frequencies instead. Nevertheless, other search procedures might be beneficial when only the *maximum a posteriori* (MAP) model and not *e.g.* the marginal posterior inclusion probabilities for the covariates are of interest.

5. Applications

We illustrate the use of the hyper- g prior for additive modelling in Section 5.1 and for logistic regression in Section 5.2.

5.1. Ozone Data

We apply our additive modelling approach to the ozone data from [Breiman and Friedman \(1985\)](#) on the association between $p = 9$ meteorological covariates and the maximum

one-hour average ozone concentration for $n = 330$ days in 1976 (see Table 1 for details). As the computational complexity of the marginal likelihood (12) is cubic in the spline basis dimension K (see Appendix A.1), we want to use splines with few, quantile-based knots. Therefore, we choose cubic O’Sullivan splines (Wand and Ormerod, 2008). Here, we get basis matrices \mathbf{Z}_j with $K = 6$ columns from 4 inner knots at the quintiles.

Variable	Description
y	Maximum 1-hour average ozone level [ppm]
x_1	Day of the year
x_2	500 millibar pressure height [m]
x_3	Wind speed [mph]
x_4	Relative humidity [%]
x_5	Temperature at Sandberg, CA [°F]
x_6	Inversion base height [feet]
x_7	Pressure gradient [mm Hg]
x_8	Visibility [miles]
x_9	Inversion base temperature [°F]

Table 1 – Description of the variables in the ozone data set.

Exhaustive evaluation of the posterior model probabilities $f(\mathbf{d} | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{d})f(\mathbf{d})$ of all $(K + 1)^9 = 40\,353\,607$ models takes 531 minutes on a standard 2.8 GHz CPU. We also applied the stochastic search algorithm described in Section 4 with 10^6 iterations, which took 2 minutes and resulted in 125 619 models. 92.4% of the posterior model probability have been discovered and the 1858 top models were found by this algorithm.

In Table 2 the marginal posterior probabilities for linear and smooth inclusion of the nine covariates are shown. While x_1 , x_5 , x_7 and x_8 are clearly included as smooth terms, there is considerable uncertainty for the other covariates whether to be included linearly or smoothly. Only the overall inclusion probability for x_6 is below 50%.

The MAP model includes smooth terms for x_1 , x_5 , x_7 , x_8 and x_9 . The covariates

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
not included ($d_j = 0$)	0.00	0.02	0.33	0.06	0.00	0.59	0.00	0.00	0.28
linear ($d_j = 1$)	0.00	0.46	0.32	0.35	0.00	0.14	0.00	0.05	0.22
smooth ($d_j > 1$)	1.00	0.52	0.35	0.59	0.99	0.28	1.00	0.95	0.50

Table 2 – Marginal posterior inclusion probabilities in the ozone data set.

x_2 and x_4 are included linearly while x_3 and x_6 are not included. Figure 1 shows the estimated covariate effects, which were obtained from 10 000 posterior samples. Note that for linear functions m_j , the pointwise credible intervals coincide with the simultaneous credible intervals (Besag, Green, Higdon, and Mengersen, 1995, p. 30). This is because all straight lines samples intersect in one point, which is $(\bar{x}_j, 0)$ due to the centring of the covariates in (4).

In comparison to the MAP model in Sabanés Bové and Held (2011b, section 4) based on Bayesian fractional polynomials, similar functional forms are estimated for the effects of x_1 , x_4 and x_5 , while differences are visible for x_7 and x_8 . Note that x_6 is included in the MAP model in Sabanés Bové and Held (2011b). See also Casella and Moreno (2006) for an objective Bayesian variable selection analysis (without the possibility of smooth effects) of this data set.

5.2. Pima Indian Diabetes Data

We now apply the generalised additive modelling approach to the logistic regression of $p = 7$ potential risk factors on the presence of diabetes in $n = 532$ women of Pima Indian heritage (Frank and Asuncion, 2010; Ripley, 1996), see Table 3 for details. As for the ozone data set, we use cubic O’Sullivan splines with inner knots at the quintiles and a uniform hyper-prior on $g/(g + 1)$, and explore the model space of dimension $7^7 = 823\,543$ with 10^6 iterations of the stochastic search algorithm. The computational complexity is higher than for the normal response case, with 218 minutes required for the evaluation of 43 766 models. We validated the results with an exhaustive evaluation of all models, requiring 94 hours. Indeed, the stochastic search found 98.3% of the posterior probability mass and the 489 top models.

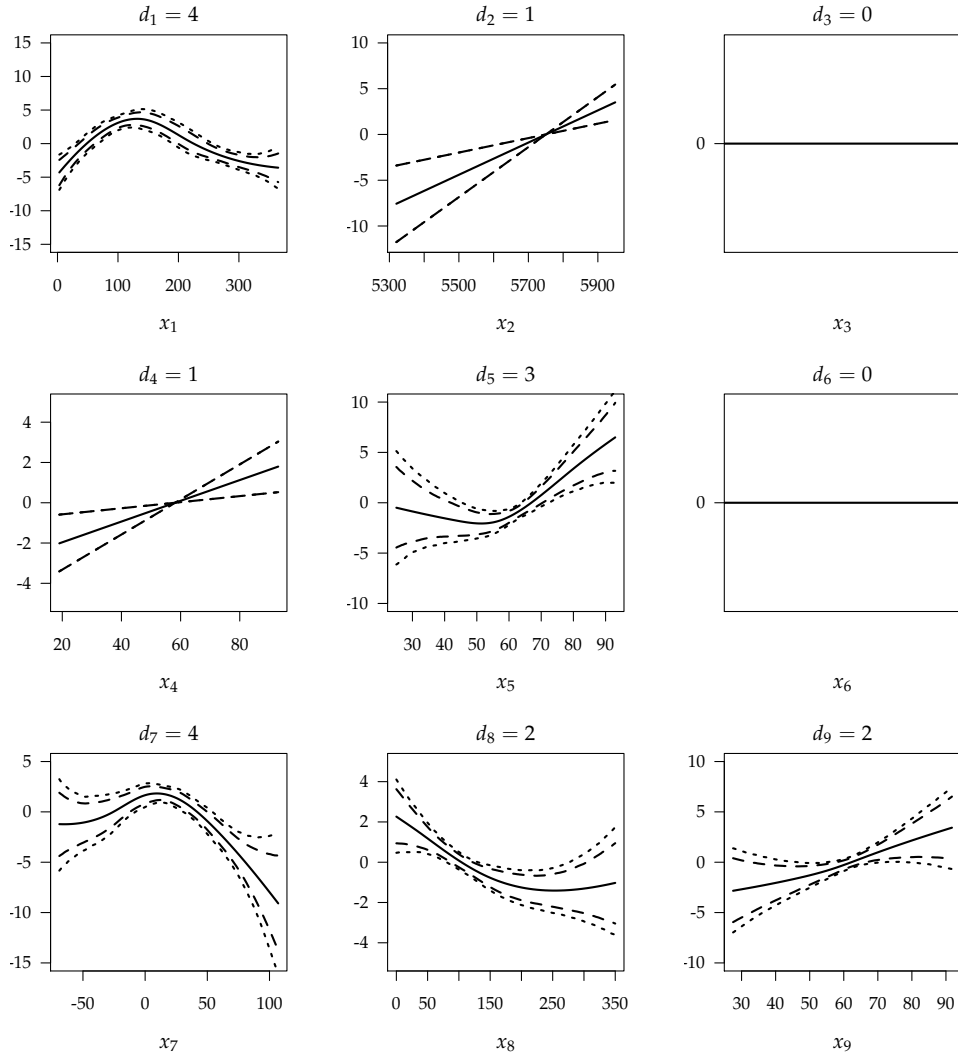


Figure 1 – Estimated covariate effects in the MAP model for the ozone data, based on 10 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.

Variable	Description
y	Signs of diabetes according to WHO criteria (Yes = 1, No = 0)
x_1	Number of pregnancies
x_2	Plasma glucose concentration in an oral glucose tolerance test [mg/dl]
x_3	Diastolic blood pressure [mm Hg]
x_4	Triceps skin fold thickness [mm]
x_5	Body mass index (BMI) [kg/m ²]
x_6	Diabetes pedigree function
x_7	Age [years]

Table 3 – Description of the variables in the Pima Indian diabetes data set.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
not included ($d_j = 0$)	0.63	0.00	0.81	0.84	0.00	0.02	0.01
linear ($d_j = 1$)	0.09	0.48	0.09	0.06	0.11	0.26	0.00
smooth ($d_j > 1$)	0.28	0.52	0.09	0.10	0.88	0.72	0.99

Table 4 – Marginal posterior inclusion probabilities in the Pima Indian diabetes data set.

In Table 4 the marginal posterior probabilities for linear and smooth inclusion of the covariates are shown. There is clear evidence for inclusion of the covariates x_2 , x_5 , x_6 and x_7 , which have posterior inclusion probabilities over 98%. For the other three covariates, the inclusion probability is below 40%. Smooth modelling of the effects of x_5 , x_6 and x_7 seems to be necessary, while this is not so clear for x_2 .

Figure 2 shows the estimated covariate effects in the MAP model which features a linear term for x_2 and smooth terms for x_5 , x_6 and x_7 . The estimates are obtained from 10 000 MCMC samples.² Note that the Chib and Jeliazkov (2001) estimate (−243.426, MCMC standard error 0.008) of the log marginal likelihood of the MAP model, which was also computed, is quite close to the integrated Laplace approximation (−243.511).

²Every 2nd sample was saved after burning the first 1000 iterations, with acceptance rate 67% using two IWLS steps per proposal.

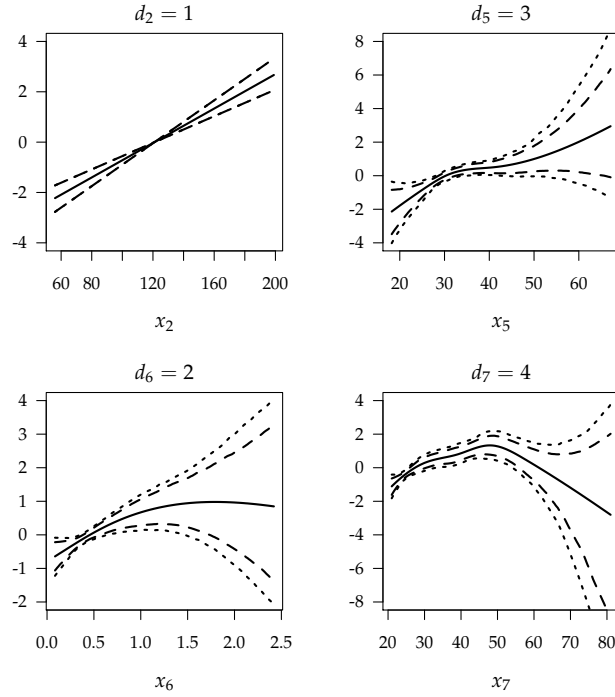


Figure 2 – Estimated covariate effects in the MAP model for the Pima Indian diabetes data set, based on 10 000 MCMC samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.

This indicates that the integrated Laplace approximation is fairly accurate.

The results are qualitatively similar to those obtained with a fractional polynomial modelling approach by [Sabanés Bové and Held \(2011a, section 5\)](#) and with a cubic smoothing spline approach by [Cottet et al. \(2008, section 3.2\)](#). It is interesting that in the earlier work by [Yau et al. \(2003, section 5.2\)](#), a very low posterior inclusion probability (0.07) for x_6 was reported for a different subset of the original Pima Indian diabetes data set. If pure variable selection without covariate transformation is considered, as in [Holmes and Held \(2006, section 2.6\)](#) and [Sabanés Bové and Held \(2011a, section 4\)](#), the strong nonlinear effect of x_7 is missed completely, and instead x_1 gets a higher posterior inclusion probability. This highlights the importance of allowing nonlinear covariate effects.

6. Postprocessing

Given the list of all possible models $\mathbf{d} \in \mathcal{D}^p$, or a subset found by the stochastic search procedure described in Section 4, one may consider postprocessing the results.

First, when the main interest lies in variable selection, the models which feature the same covariates can be summarised into a meta-model: The posterior probabilities of the sub-models are summed up to give the posterior probability of the meta-model, and estimates in the meta-model are obtained by averaging the sub-models with weights proportional to their posterior probabilities (see *e.g.* [Hoeting, Madigan, Raftery, and Volinsky, 1999](#), for model averaging). For example, the best meta-model for the ozone data features all covariates except x_6 and has posterior probability 0.261. The corresponding estimates of the covariate effects are shown in Figure 3. For the Pima Indian diabetes data the meta-model with highest posterior probability (0.466) includes x_2 , x_5 , x_6 and x_7 . In both examples, the best meta-model happens to be identical with the median probability meta-model, which features all covariates having marginal posterior inclusion probability greater than 50% ([Barbieri and Berger, 2004](#)), *cp.* Tables 2 and 4. Similarly, it could be interesting to summarise models which only differ in the degrees of freedom for smooth terms. This would correspond to the situation of testing linearity versus nonlinearity of covariate effects (*cp.* Section 4).

Second, in order to allow for continuous degrees of freedom, one can optimise the marginal likelihood of the MAP model with respect to the degrees of freedom of the covariates included. That is, an optimisation of $f(\mathbf{y} | \mathbf{d})$ over the *continuous* range $(1, K + 1)$ is performed for all d_j 's with $d_j > 0$ in the MAP model. For example, the MAP configuration for the ozone data is $(4, 1, 0, 1, 3, 0, 4, 2, 2)$ and the resulting optimised configuration is $(4.35, 1, 0, 1.08, 3.44, 0, 3.63, 2.3, 1)$, which increases the log marginal likelihood from -1413.86 to -1412.91 . Although d_9 decreases from 2 to 1 (rounded down to 2 decimals), the function estimates are very similar to those from the MAP model in Figure 1. For the Pima Indian diabetes data, the log marginal likelihood increases from -243.51 for the MAP model $(0, 1, 0, 0, 3, 2, 4)$ to -243.39 for the optimised configuration $(0, 1, 0, 0, 3.36, 2.09, 3.73)$. Again the function estimates are omitted because they are close to those from the MAP model in Figure 2.

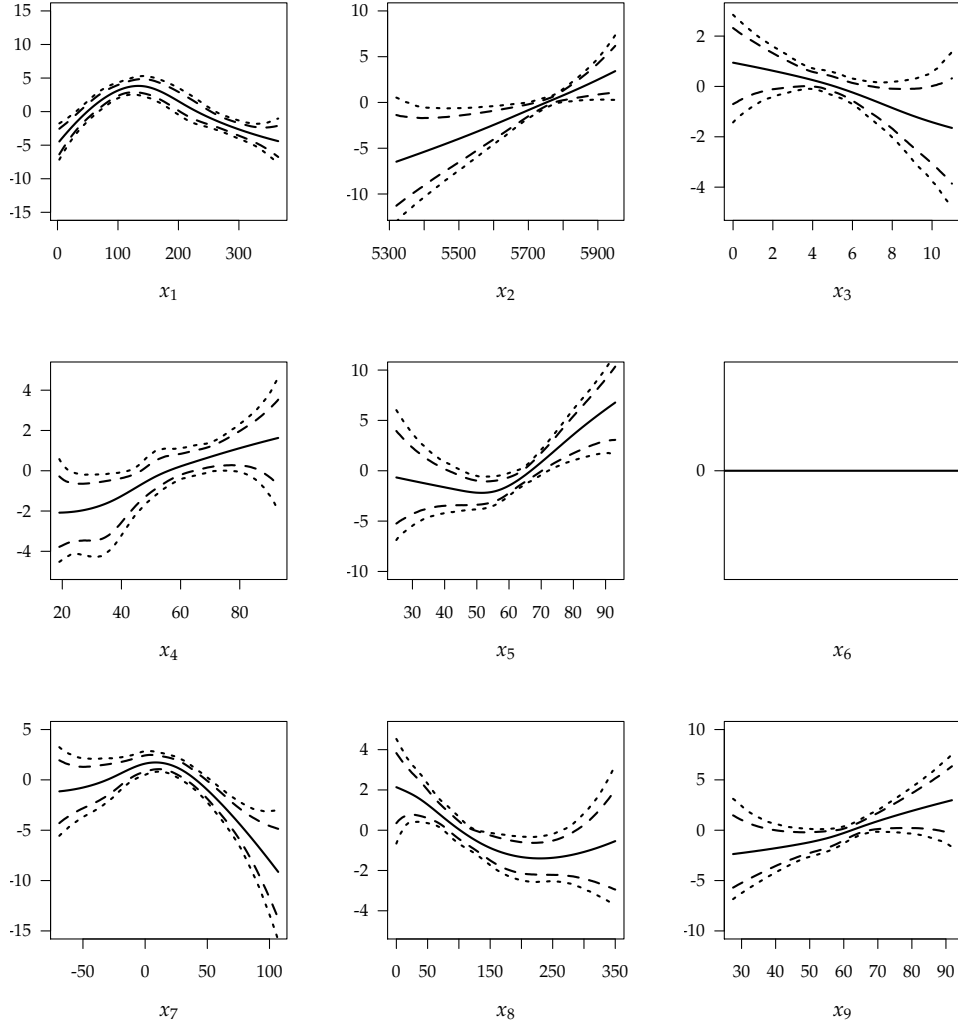


Figure 3 – Estimated covariate effects in the best meta-model (and median probability meta-model) for the ozone data, based on 20 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.

7. Discussion

Our Bayesian approach to simultaneous variable and function selection in generalised regression is based on fixed-dimension spline bases and penalty-parameter smoothness control. In this respect, it differs from knot-selection approaches such as [Smith and Kohn \(1996\)](#) and [Denison, Mallick, and Smith \(1998\)](#). We are only considering penalties on a fixed grid of values, which scales automatically for each covariate via the degrees of freedom transformation. In this regard, our approach is close to many popular Lasso-type proposals, which optimise the tuning-parameters on a fixed grid via cross-validation (*e.g.* [Zou and Hastie, 2005](#)). [Cantoni and Hastie \(2002\)](#) propose a likelihood-ratio-type test statistic to compare additive models with different degrees of freedom. [Fong, Rue, and Wakefield \(2010\)](#) use a similar scaling to examine the prior on the degrees of freedom implied by the prior on the variance component in a generalised linear mixed model. They also use O’Sullivan spline bases as we did in our examples, but they do not consider variable selection.

In a frequentist setting, [Marra and Wood \(2011, section 2.1\)](#) propose to use an additional penalty on the linear part of the spline function in order to shrink it adaptively to zero. To include variable selection, a lower threshold for the effective degrees of freedom must be chosen. Our generalised g -prior [\(27\)](#) also shrinks the linear parts of the spline functions to zero, where the prior covariance matrix takes the correlations between the covariates into account. Furthermore, we explicitly ex- or include covariates and then compare the resulting models based on their posterior probabilities.

We propose a conventional prior for the intercept and the linear coefficients, which directly generalises the hyper- g prior in the linear model ([Liang et al., 2008](#)) and in the generalised linear model ([Sabanés Bové and Held, 2011a](#)). [Pauler \(1998\)](#) proposes a related unit-information prior for the fixed effects in linear mixed models, but fixes $g = n$ in [\(10\)](#). [Overstall and Forster \(2010\)](#) propose a unit-information prior for the fixed effects in generalised linear mixed models, but the information matrix is based on the first-stage likelihood and not on the integrated likelihood as in our approach. Also, no hyper-prior on the parameter g is considered, because it is fixed at $g = n$. As they use an inverse-Wishart prior on the covariance matrix of the random effects, their approach is perhaps better suited to generic random effects models. [Forster, Gill, and Overstall \(2010\)](#) propose a novel reversible-jump MCMC algorithm to infer posterior model

probabilities.

In future work, we would like to combine the semiparametric splines with the parametric fractional polynomials (Sabanés Bové and Held, 2011b). The idea is that a smooth term $m_j(x_j)$ could also be modelled by a fractional polynomial instead of a spline. This extension could be implemented coherently, because the prior formulations are compatible. With such a general framework, the important question whether a parsimonious fractional polynomial (*e.g.* $m_7 = x_7\beta_{71} + x_7^2\beta_{72}$ in the Pima Indian diabetes data example) is sufficient could be answered via posterior probabilities (see Strasak, Umlauf, Pfeiffer, and Lang (2011) for a simulation study comparing the step-wise fractional polynomial approach by Royston and Sauerbrei (2008) with penalised spline approaches).

A. Implementation details

Section A.1 gives details on the computation of the marginal likelihood (12) for normal additive models. Section A.2 derives the proposal probabilities for the stochastic search described in Section 4.

A.1. Marginal likelihood computation

A Cholesky factorisation of the covariance matrix $V_d \in \mathbb{R}^{n \times n}$ has complexity $\mathcal{O}(n^3)$ and is therefore computationally expensive. Therefore, it is advisable to avoid it and instead work with the formula

$$V_d^{-1} = I_n - Z_d M_d^{-1} Z_d^T$$

for the precision matrix, where $M_d = Z_d^T Z_d + D_d^{-1}$. The latter matrix has dimension JK , which is usually smaller than n , provided the spline basis dimension K is small. Thus, the Cholesky factorisation $M_d = M_d^{T/2} M_d^{1/2}$ is relatively fast, and we compute $W_d = Z_d M_d^{-1/2}$ such that $V_d^{-1} = I_n - W_d W_d^T$.

For the coefficient of determination $\tilde{R}_d^2 = SSM_d / SST_d$, we need to compute the sum of squares in total (SST_d) and the sum of squares explained by the model (SSM_d). For

SST_d , we have

$$\begin{aligned} SST_d &= (\mathbf{y} - \mathbf{1}_n \bar{y})^T \mathbf{V}_d^{-1} (\mathbf{y} - \mathbf{1}_n \bar{y}) \\ &= \|\mathbf{y} - \mathbf{1}_n \bar{y}\|^2 - \|\mathbf{W}_d^T (\mathbf{y} - \mathbf{1}_n \bar{y})\|^2. \end{aligned}$$

Note that the first term in (12) can be written as $\left\| \mathbf{V}_d^{-T/2} (\mathbf{y} - \mathbf{1}_n \bar{y}) \right\|^{-(n-1)} = SST_d^{-(n-1)/2}$. For SSM_d , note that the fit of the general linear model is $\hat{\mathbf{y}}_d = \mathbf{1}_n \bar{y} + \mathbf{X}_d \hat{\boldsymbol{\beta}}_d$, where

$$\hat{\boldsymbol{\beta}}_d = (\mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d)^{-1} \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{y}$$

is the weighted least squares estimate of $\boldsymbol{\beta}_d$. Therefore

$$\begin{aligned} SSM_d &= (\hat{\mathbf{y}}_d - \mathbf{1}_n \bar{y})^T \mathbf{V}_d^{-1} (\hat{\mathbf{y}}_d - \mathbf{1}_n \bar{y}) \\ &= \hat{\boldsymbol{\beta}}_d^T \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \hat{\boldsymbol{\beta}}_d \end{aligned}$$

can be computed by Cholesky factorising $\mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d = \mathbf{C}_d^T \mathbf{C}_d$, solving the triangular system $\mathbf{C}_d^T \mathbf{v}_d = \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{y}$ and setting $SSM_d = \|\mathbf{v}_d\|^2$.

Finally, to compute the determinant term in (12), we can again avoid factorising \mathbf{V}_d , because we have

$$\left| \mathbf{V}_d^{1/2} \right|^{-1} = \left| \mathbf{V}_d^{-1} \right|^{1/2} = \left| \mathbf{I}_n - \mathbf{W}_d \mathbf{W}_d^T \right|^{1/2} = \left| \mathbf{I}_{JK} - \mathbf{W}_d^T \mathbf{W}_d \right|^{1/2},$$

see Harville (1997, p. 416) for the last equation. So again only a matrix of dimension JK , namely $\mathbf{I}_{JK} - \mathbf{W}_d^T \mathbf{W}_d$, needs to be factorised. Here, a LU factorisation can be used.

A.2. Proposal probabilities

First note that the two proposal types ‘Move’ and ‘Swap’ do not overlap, because a ‘Move’ always changes exactly one d_j , while a ‘Swap’ either changes none or two d_j ’s. Denote with p_m the probability to choose a ‘Move’.

Suppose a ‘Move’ was proposed for covariate $j \in \{0, 1, \dots, p\}$. We then have

$$q(\mathbf{d}' | \mathbf{d}) = p_m \cdot \frac{1}{p} \cdot \begin{cases} 1, & d_j \in \{0, K\}, \\ \frac{1}{2}, & \text{else} \end{cases}$$

and analogously

$$q(\mathbf{d} | \mathbf{d}') = p_m \cdot \frac{1}{p} \cdot \begin{cases} 1, & d'_j \in \{0, K\}, \\ \frac{1}{2}, & \text{else} \end{cases}$$

with proposal ratio

$$\frac{q(\mathbf{d}' | \mathbf{d})}{q(\mathbf{d} | \mathbf{d}')} = \begin{cases} \frac{1}{2}, & d'_j \in \{0, K\}, \\ 2, & d_j \in \{0, K\}, \\ 1, & \text{else.} \end{cases}$$

For the ‘Swap’ proposal, suppose covariates i and j are proposed to interchange their model parameters d_i and d_j . Of course, if $d_i = d_j$, then the proposal ratio equals unity because $\mathbf{d}' = \mathbf{d}$. In the other case, both model parameters are changed, and

$$q(\mathbf{d}' | \mathbf{d}) = q(\mathbf{d} | \mathbf{d}') = (1 - p_m) \cdot \binom{p}{2}^{-1},$$

so that for a ‘Swap’ we always have $q(\mathbf{d}' | \mathbf{d}) / q(\mathbf{d} | \mathbf{d}') = 1$.

B. Approximate Fisher Information for non-normal response

In this section, we present a formal derivation of (28) as the approximate Fisher information obtained from a Laplace approximation to $f(\mathbf{y} | \beta_0, \beta_d)$. For ease of notation we restrict the presentation to canonical response functions where $\eta = \theta$ and omit subscripts where they are not necessary for understanding. With $\Phi = \text{diag}\{\phi/w_i\}_{i=1}^n$, we can then rewrite the likelihood (19) as

$$f(\mathbf{y} | \beta_0, \beta, \mathbf{u}) \propto \exp \left\{ \mathbf{y}^T \Phi^{-1} \boldsymbol{\eta} - \mathbf{1}^T \Phi^{-1} b(\boldsymbol{\eta}) \right\}. \quad (32)$$

We will now use the Laplace approximation to integrate (32) over \mathbf{u} with respect to the prior $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$. The Laplace approximation has proven to be useful in the generalised linear mixed model framework, see *e.g.* Breslow and Clayton (1993). Specifically, the use of the Laplace approximation for penalised spline smoothing has been investigated by Kauermann, Krivobokova, and Fahrmeir (2009), who prove its asymptotic validity.

For the Laplace approximation of $f(\mathbf{y} | \beta_0, \beta)$ we first need to maximise the unnormalised log posterior of \mathbf{u} ,

$$\begin{aligned} l(\mathbf{u}) &= \log f(\mathbf{y} | \beta_0, \beta, \mathbf{u}) + \log f(\mathbf{u}) \\ &= \mathbf{y}^T \Phi^{-1} \boldsymbol{\eta} - \mathbf{1}^T \Phi^{-1} b(\boldsymbol{\eta}) - \frac{1}{2} \mathbf{u}^T \mathbf{D}^{-1} \mathbf{u} + \text{const}, \end{aligned} \quad (33)$$

where β_0 and β in $\eta = \mathbf{1}\beta_0 + \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$ are considered to be fixed. The corresponding score vector is

$$\begin{aligned}\frac{d}{d\mathbf{u}}l(\mathbf{u}) &= \mathbf{Z}^T \Phi^{-1} \mathbf{y} - \mathbf{Z}^T \text{diag}\{b'(\eta)\} \Phi^{-1} \mathbf{1} - D^{-1} \mathbf{u} \\ &= \mathbf{Z}^T \Phi^{-1} (\mathbf{y} - \boldsymbol{\mu}) - D^{-1} \mathbf{u},\end{aligned}$$

where $\boldsymbol{\mu} = b'(\eta)$, and the corresponding Hessian is

$$\begin{aligned}\frac{d}{d\mathbf{u}} \frac{d}{d\mathbf{u}^T} l(\mathbf{u}) &= \frac{d}{d\mathbf{u}} \left\{ (\mathbf{y} - \boldsymbol{\mu})^T \Phi^{-1} \mathbf{Z} - \mathbf{u}^T D^{-1} \right\} \\ &= -\mathbf{Z}^T \mathbf{W}(\eta) \mathbf{Z} - D^{-1}.\end{aligned}$$

Making one Newton-Raphson step from the starting point $\mathbf{u} = \mathbf{0}$, we get the approximate mode \mathbf{u}^* of $l(\mathbf{u})$:

$$\begin{aligned}\mathbf{u}^* &= \mathbf{0} - \left(\frac{d}{d\mathbf{u}} \frac{d}{d\mathbf{u}^T} l(\mathbf{0}) \right)^{-1} \frac{d}{d\mathbf{u}} l(\mathbf{0}) \\ &= \left(\mathbf{Z}^T \mathbf{W}(\eta_L) \mathbf{Z} + D^{-1} \right)^{-1} \mathbf{Z}^T \Phi^{-1} (\mathbf{y} - \boldsymbol{\mu}_L),\end{aligned}\tag{34}$$

where $\eta_L = \mathbf{1}\beta_0 + \mathbf{X}\beta$ and $\boldsymbol{\mu}_L = b'(\eta_L)$. Note that this corresponds to the result of a second-order Taylor expansion of $l(\mathbf{u})$ around $\mathbf{u} = \mathbf{0}$. Hence, the Laplace approximation of $f(\mathbf{y} | \beta_0, \beta)$ is

$$\begin{aligned}\tilde{f}(\mathbf{y} | \beta_0, \beta) &\propto \exp(l(\mathbf{u}^*)) (2\pi)^{JK/2} \left| -\frac{d}{d\mathbf{u}} \frac{d}{d\mathbf{u}^T} l(\mathbf{u}^*) \right|^{-1/2} \\ &= \exp \left(\mathbf{y}^T \Phi^{-1} \eta^* - \mathbf{1}^T \Phi^{-1} b(\eta^*) - \frac{1}{2} \mathbf{u}^{*T} D^{-1} \mathbf{u}^* \right) \\ &\quad \times (2\pi)^{JK/2} \left| \mathbf{Z}^T \mathbf{W}(\eta^*) \mathbf{Z} + D^{-1} \right|^{-1/2},\end{aligned}\tag{35}$$

where JK is the dimension of \mathbf{u} .

In order to derive the approximate Fisher information of β from $\tilde{f}(\mathbf{y} | \beta_0, \beta)$, we make two additional simplifying assumptions: First, we assume that $\mathbf{W}(\eta)$ does not vary much in β , so that we can ignore the determinant in (35), for example. This is a common simplification, suggested *e.g.* in [Breslow and Clayton \(1993\)](#). Second, we approximate $b(\eta^*)$ by a second-order Taylor expansion of $b(\eta)$ around η_L , yielding

$$\mathbf{1}^T \Phi^{-1} b(\eta^*) \approx \mathbf{1}^T \Phi^{-1} b(\eta_L) + \boldsymbol{\mu}_L^T \Phi^{-1} \mathbf{Z} \mathbf{u}^* + \frac{1}{2} \mathbf{u}^{*T} \mathbf{Z}^T \mathbf{W}_L \mathbf{Z} \mathbf{u}^*,$$

where $W_L = W(\eta_L)$. Using these two simplifications and plugging in (34), we arrive at the expression

$$\begin{aligned}\log \tilde{f}(\mathbf{y} | \beta_0, \beta) &= \mathbf{y}^T \Phi^{-1} \eta_L - \mathbf{1}^T \Phi^{-1} b(\eta_L) \\ &\quad + (\mathbf{y} - \mu_L)^T \Phi^{-1} \mathbf{Z} u^* - \frac{1}{2} u^{*T} (\mathbf{Z}^T W_L \mathbf{Z} + D^{-1}) u^* \\ &= \mathbf{y}^T \Phi^{-1} \eta_L - \mathbf{1}^T \Phi^{-1} b(\eta_L) \\ &\quad + \frac{1}{2} (\mathbf{y} - \mu_L)^T \Phi^{-1} \mathbf{Z} (\mathbf{Z}^T W_L \mathbf{Z} + D^{-1})^{-1} \mathbf{Z}^T \Phi^{-1} (\mathbf{y} - \mu_L)\end{aligned}\quad (36)$$

for the approximate marginal log-likelihood of β_0 and β . From (36) we can finally approximate the Fisher information $J(\beta_0, \beta) = -\frac{d}{d\beta} \frac{d}{d\beta^T} \log f(\mathbf{y} | \beta_0, \beta)$ as

$$\begin{aligned}\tilde{J}(\beta_0, \beta) &= -\frac{d}{d\beta} \frac{d}{d\beta^T} \log \tilde{f}(\mathbf{y} | \beta_0, \beta) \\ &= \mathbf{X}^T W_L^{1/2} \left(\mathbf{I} - W_L^{1/2} \mathbf{Z} (\mathbf{Z}^T W_L \mathbf{Z} + D^{-1})^{-1} \mathbf{Z}^T W_L^{1/2} \right) W_L^{1/2} \mathbf{X} \quad (37) \\ &= \mathbf{X}^T W_L^{1/2} (\mathbf{I} + W_L^{1/2} \mathbf{Z} D \mathbf{Z}^T W_L^{1/2})^{-1} W_L^{1/2} \mathbf{X}.\end{aligned}\quad (38)$$

Evaluating the approximate Fisher information at $\beta_0 = 0, \beta = \mathbf{0}$, such that $W_L = W(\mathbf{0})$, we recognise that $\tilde{J}(0, \mathbf{0})$ from (38) is identical to J_0 in (28). Note that the representation (37) can be better suited for computation: the first paragraph of Appendix A.1 applies here after replacing \mathbf{Z}_d with $W_0^{1/2} \mathbf{Z}$.

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth Dover printing, tenth GPO printing edition, 1964. ISBN 0-486-61272-4.

M. Aerts, G. Claeskens, and M. P. Wand. Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, 103(1-2):455–470, 2002. ISSN 0378-3758. URL <http://www.sciencedirect.com/science/article/B6V0M-45DDDTTC-11/2/90b0f85072547a973fda99>

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004. ISSN 0090-5364. doi: 10.1214/009053604000000238.

- C. Belitz and S. Lang. Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis*, 53(1):61–81, 2008. ISSN 0167-9473. URL <http://www.sciencedirect.com/science/article/B6V8V-4SRW0YY-1/2/e42416ad784917ca749d51f>.
- J. O. Berger and L. R. Pericchi. Objective Bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series*, 38(1):135–207, 2001. ISSN 07492170. URL <http://www.jstor.org/stable/4356165>.
- J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147, 1979. ISSN 00359246. URL <http://www.jstor.org/stable/2985028>.
- J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems (with discussion). *Statistical Science*, 10(1):3–66, 1995. URL citeseer.ist.psu.edu/besag95bayesian.html.
- Å. Björck. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT Numerical Mathematics*, 7(1):1–21, 1967. ISSN 0006-3835. URL <http://dx.doi.org/10.1007/BF01934122>.
- L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.
- N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993. ISSN 01621459. URL <http://www.jstor.org/stable/2290687>.
- A. Brezger and S. Lang. Simultaneous probability statements for Bayesian P-splines. *Statistical Modelling*, 8(2):141–168, 2008. ISSN 1471082X. URL <http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=34389786&site=ehost-live>.
- E. Cantoni and T. Hastie. Degrees-of-freedom tests for smoothing splines. *Biometrika*, 89(2):251–263, 2002. ISSN 00063444. URL <http://www.jstor.org/stable/4140575>.
- G. Casella and E. Moreno. Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167, 2006.

- S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001. ISSN 01621459. URL <http://www.jstor.org/stable/2670365>.
- M. A. Clyde and J. Ghosh. A note on the bias in estimating posterior probabilities in variable selection. Technical report, Duke University, 2010. URL <ftp.stat.duke.edu/WorkingPapers/10-11.pdf>.
- M. A. Clyde, J. Ghosh, and M. L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101, 2011. doi: 10.1198/jcgs.2010.09049. URL <http://pubs.amstat.org/doi/pdfplus/10.1198/jcgs.2010.09049>.
- R. Cottet, R. J. Kohn, and D. J. Nott. Variable selection and model averaging in semiparametric overdispersed generalized linear models. *Journal of the American Statistical Association*, 103(482):661–671, 2008. ISSN 0162-1459. URL <http://dx.doi.org/10.1198/016214508000000346>.
- W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008. ISSN 0378-3758. URL <http://www.sciencedirect.com/science/article/B6V0M-4NK4G63-2/2/e2c24b01d141b02fe783c08>.
- D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 60(2):333–350, 1998. ISSN 13697412. URL <http://www.jstor.org/stable/2985943>.
- P. H. C. Eilers and B. D. Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews Computational Statistics*, 2(6):637–653, 2010. ISSN 1939-0068. URL <http://dx.doi.org/10.1002/wics.125>.
- L. Fahrmeir, T. Kneib, and S. Lang. Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, 14(3):715–745, 2004.
- L. Fahrmeir, T. Kneib, and S. Konrath. Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20(2):203–219, 2010. ISSN 0960-3174. URL <http://dx.doi.org/10.1007/s11222-009-9158-3>.

- Y. Fong, H. Rue, and J. Wakefield. Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412, 2010. URL <http://biostatistics.oxfordjournals.org/cgi/content/abstract/11/3/397>.
- J. Forster, R. Gill, and A. Overstall. Reversible jump methods for generalised linear models and generalised linear mixed models. *Statistics and Computing*, 2010. ISSN 0960-3174. URL <http://dx.doi.org/10.1007/s11222-010-9210-3>. epub ahead of print.
- A. Forte. *Objective Bayes Criteria for Variable Selection*. PhD thesis, Universitat de València, 2011.
- A. Frank and A. Asuncion. UCI Machine Learning Repository, 2010. URL <http://archive.ics.uci.edu/ml>.
- J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL <http://www.jstor.org/stable/2699986>.
- G. García-Donato and M. A. Martínez-Beneito. Inferences in Bayesian variable selection problems with large model spaces. Technical report, Universidad de Castilla La Mancha, Spain, 2011. URL <http://arxiv.org/abs/1101.4368>.
- C. Hans, A. Dobra, and M. West. Shotgun stochastic search for “large p” regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.
- D. A. Harville. *Matrix Algebra From a Statistician’s Perspective*. Springer, New York, 1997.
- T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.
- L. Held. Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, 13(1):20–35, 2004. ISSN 10618600. URL <http://www.jstor.org/stable/1391142>.
- H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60, 1981. ISSN 00361445. URL <http://www.jstor.org/stable/2029838>.

- J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999. URL <http://www.stat.washington.edu/www/research/online/hoeting1999.pdf>.
- C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.
- G. Kauermann and G. Tutz. Testing generalized linear and semiparametric models against smooth alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 63(1):147–166, 2001. ISSN 1467-9868. URL <http://dx.doi.org/10.1111/1467-9868.00281>.
- G. Kauermann, T. Krivobokova, and L. Fahrmeir. Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(2):487–503, 2009. ISSN 13697412. URL <http://www.jstor.org/stable/40247584>.
- T. Kneib, T. Hothorn, and G. Tutz. Variable selection and model choice in geoadaptive regression models. *Biometrics*, 65(2):626–634, 2009. ISSN 1541-0420. URL <http://dx.doi.org/10.1111/j.1541-0420.2008.01112.x>.
- F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.
- D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232, 1995.
- G. Marra and S. N. Wood. Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, 55(7):2372–2387, 2011. ISSN 0167-9473. URL <http://www.sciencedirect.com/science/article/B6V8V-524WDXR-2/2/a61fbfea37502858cfa0351>.
- P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman and Hall, New York, second edition, 1989.

- L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Annals of Statistics*, 37(6B):3779–3821, 2009.
- A. M. Overstall and J. J. Forster. Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics and Data Analysis*, 54(12):3269–3288, 2010. ISSN 0167-9473. doi: 10.1016/j.csda.2010.03.008. URL <http://www.sciencedirect.com/science/article/B6V8V-4YP8TGG-2/2/17f2c7cb3e2907a5326ca6c>.
- A. Panagiotelis and M. Smith. Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics*, 143(2):291–316, 2008. ISSN 0304-4076. URL <http://www.sciencedirect.com/science/article/B6VC0-4R17V3R-1/2/7774139176daf9ebb0b8c7f>.
- D. K. Pauler. The Schwarz criterion and related methods for normal linear models. *Biometrika*, 85(1):13–27, 1998. URL <http://biomet.oxfordjournals.org/content/85/1/13.abstract>.
- P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: Sparse additive models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1201–1208, Cambridge, MA, 2008. MIT Press.
- B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.
- P. Royston and W. Sauerbrei. *Multivariable Model-building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables*. Wiley Series in Probability and Statistics. Wiley, Chichester, 2008. URL <http://www.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/>.
- H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(2):319–392, 2009. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2008.00700.x.
- D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2003.

- D. Ruppert, M. Wand, and R. Carroll. Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, 3(1):1193–1256, 2009.
- D. Sabanés Bové and L. Held. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6, 2011a. URL <http://ba.stat.cmu.edu/abstracts/Sabanés.php>. Forthcoming article as of 18/2/2011.
- D. Sabanés Bové and L. Held. Bayesian fractional polynomials. *Statistics and Computing*, 21(3):309–324, 2011b. doi: 10.1007/s11222-010-9170-7. URL <http://dx.doi.org/10.1007/s11222-010-9170-7>.
- F. Scheipl. Normal-mixture-of-inverse-gamma priors for Bayesian regularization and model selection in structured additive regression models. Technical Report 84, Department of Statistics, University of Munich, 2010. URL <http://epub.ub.uni-muenchen.de/11785/>.
- J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619, 2010.
- M. Smith and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996. ISSN 0304-4076. URL <http://www.sciencedirect.com/science/article/B6VC0-3VWT1X6-X/2/e4dc8a22271b240b3768be8>.
- A. M. Strasak, N. Umlauf, R. M. Pfeiffer, and S. Lang. Comparing penalized splines and fractional polynomials for flexible modelling of the effects of continuous predictor variables. *Computational Statistics and Data Analysis*, 55(4):1540–1551, 2011. ISSN 0167-9473. URL <http://www.sciencedirect.com/science/article/B6V8V-51920YH-2/2/c419419d850733524615d36>.
- G. Tutz and H. Binder. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62(4):961–971, 2006. ISSN 1541-0420. URL <http://dx.doi.org/10.1111/j.1541-0420.2006.00578.x>.
- M. P. Wand and J. T. Ormerod. On semiparametric regression with O’Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50(2):179–198, 2008. ISSN 1467-842X. URL <http://dx.doi.org/10.1111/j.1467-842X.2008.00507.x>.

- M. West. Generalized linear models: scale parameters, outlier accommodation and prior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, pages 531–558, Amsterdam, 1985. North-Holland.
- S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, 2006.
- S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 73(1):3–36, 2011. ISSN 1467-9868. URL <http://dx.doi.org/10.1111/j.1467-9868.2010.00749.x>.
- P. Yau, R. J. Kohn, and S. Wood. Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, 12(1):23–54, 2003. ISSN 1061-8600. URL <http://dx.doi.org/10.1198/1061860031301>.
- A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, chapter 5, pages 233–243. North-Holland, Amsterdam, 1986.
- A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, pages 585–603, Valencia, 1980. University of Valencia Press.
- H. H. Zhang and Y. Lin. Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, 16(3):1021–1041, 2006.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67(2):301–320, 2005. ISSN 13697412. URL <http://www.jstor.org/stable/3647580>.